

结合网络拓扑与节点内容的统一化 半监督社团检测方法

许伟忠¹, 曹金鑫¹, 金 弟², 孙 翔³,
张晓峰¹, 刘 路³, 丁卫平¹

(1.南通大学信息科学技术学院,江苏 南通 226019)

(2.天津大学智能与计算学部,天津 300350)

(3.莱斯特大学信息学院 莱斯特,英国 LE1 7RH)

[摘要] 在复杂网络分析中,社团检测发挥着越来越重要的作用,而在实际应用中如何提高社团检测的性能仍是一个共同研究目标. 由于网络节点中内容信息有助于社团识别,一些方法侧重于将网络拓扑和节点内容相结合,并且获得了不错效果. 此外,也有些方法借用节点之间的拓扑相似度,以提升实现社团检测性能. 鉴于此,我们提出了一个统一化方法,结合节点内容的半监督社团检测,简称 SCDNC. 在该方法中,我们不仅将链接增强应用于社团检测,而且实现了拓扑和内容有机融合. 首先,我们运用随机模型来描述节点社团隶属度. 其次,我们构建出一个刻画节点内容社团隶属度的随机块模型,节点社团隶属度作为节点内容的权重向量,以实现拓扑和内容结合. 再次,我们利用网络中节点之间的拓扑相似度构建先验信息,即使网络中节点与其最相似的邻居节点具有相同的隶属度分布. 最后,使用非负矩阵分解的方法学习新模型的统一化参数. 在带有真实标签的人工网络和真实网络上,我们对新方法与一些当前流行的社团检测方法进行了性能比较. 实验结果显示,通过融合节点内容和先验信息强化的链接,新方法检测社团的性能取得了显著提升.

[关键词] 社团检测,节点内容,先验信息,随机块,非负矩阵分解

[中图分类号] TP182 [文献标志码] A [文章编号] 1001-4616(2023)01-0130-09

A Unified Semi-supervised Community Detection Approach Integrating Network Topology and Node Contents

Xu Weizhong¹, Cao Jinxin¹, Jin Di², Sun Xiang³,
Zhang Xiaofeng¹, Liu Lu³, Ding Weiping¹

(1.School of Information Science and Technology,Nantong University,Nantong 226019,China)

(2.College of Intelligence and Computing,Tianjin University,Tianjin 300350,China)

(3.School of Informatics,University of Leicester,Leicester,UK LE1 7RH)

Abstract: Community detection plays an increasing important role in complex network analysis. There is still a goal that how to improve the performance of community detection in real applications. Due to the content in networks helpful to identifying communities, some methods focus on combining network topology with node content, which obtains no bad performance of community detection. Besides, some community detection enhancement methods are mainly based on designing the topological similarity of nodes to adjust network topology, which aims to achieve the enhancement. In order to further improve the quality of community detection, we propose a unified method, Semi-supervised Community Detection with Node Contents, shorted as SCDNC, which not only apply the enhancement into community detection, but also achieve the integration of network topology and node content. In the new method, firstly, we propose a stochastic block model to describe the community memberships of nodes. Secondly, we present another stochastic model to describe

收稿日期:2021-12-31.

基金项目:国家自然科学基金面上项目(61976120)、江苏省自然科学基金面上项目(BK20191445)、江苏省高等学校自然科学研究面上项目(21KJB520018)、南通大学人才引进项目(03081198).

通讯作者:曹金鑫,博士,讲师,研究方向:数据挖掘、机器学习、社团检测等. E-mail:alfred7c@ntu.edu.cn

the community memberships of node contents, which utilizes community memberships of nodes as weight vectors of node contents. By now, integrating network topology with node content is achieved. Thirdly, we calculate the topological similarity of nodes by using links, and then model the prior information based on topological similarity, i.e., we make nodes and their most similar neighbors have the same community membership. Finally, we present a nonnegative matrix factorization approach to obtain the parameters of the model. On both synthetic and real-world networks with ground-truths, we compare performance of the new method with the state-of-the-art methods. The experimental results show that the new method obtains significant improvement for community detection via combining node contents and network topology enhanced by prior information.

Key words: community detection, node contents, prior information, stochastic block, nonnegative matrix factorization

社交网络的发展催生了纷繁复杂的海量数据,如用户关系网络、产品评论和在线评论等. 这些海量数据通常被建模为复杂网络. 对于复杂网络的分析,挖掘网络中的社团结构是一项十分关键的任务,可用以揭示复杂网络的结构和功能特性^[1].

一般来说,社团为网络中一组节点所构成的簇,同一簇的节点之间链接稠密,而不同簇的节点之间链接稀疏. 已经有不少研究人员^[2-7]提出了许多不同类型的方法以检测网络中社团结构. 这些方法很大程度上依赖于网络拓扑,其性能往往受到原始网络中某些链接丢失的影响. 实际上,真实世界中的网络充满各种类型的内容信息,比如社交网络中用户节点上的用户信息等. 这些内容能够减缓链接丢失的负面影响,辅助网络拓扑进行社团检测. Newman 等^[8]研究发现,内容信息有助于挖掘网络中的社团结构. 已有一些融合拓扑和节点内容的社团检测方法^[9-12]被提出,并且也取得了不错的效果. 但是,这些方法仍受网络拓扑中结构信息的影响.

此外,传统社团检测方法大多基于以下想法建模网络拓扑,即视社团为一组具有相似链接模式的节点集合,其在具有清晰社团结构的网络上表现效果不错. 而当网络中包含一些复杂结构时,例如同配结构、异配结构^[13]和分层结构^[14]等,这些方法的性能将急剧下降. 为了解决该问题,可挖掘网络中先验信息对网络拓扑进行调整,进而提高社团检测方法的性能. 最近, Yang 等^[15]、He 等^[16]使用先验信息提出半监督社团检测方法. 其研究结果显示,尤其是在具有噪声和复杂结构的真实网络中,这些方法仅融入有限的先验信息便实现了社团检测精度和鲁棒性的显著提升.

综上所述,网络中包含了网络拓扑、节点内容和先验信息. 当融合来自网络中不同源的有益信息,社团检测的性能将得到进一步提升. 例如,我们在网络中发现对于具有相似内容信息的节点或是邻居相同的比例高的节点,可运用节点内容、先验等有益信息指导设计社团检测算法,将节点划分到同一社团. 在本文,我们提出了一种同时融合网络拓扑、节点内容的统一化半监督社团检测模型 (semi-supervised community detection with node contents, SCDNC). 新模型的输入为两个源,其一是表示网络拓扑信息的邻接矩阵,其二是表示节点内容的特征矩阵. 此外,利用邻接矩阵转化为节点结构相似度矩阵获得先验信息. 在该模型框架下,我们首先基于生成模型的思想构建节点隶属度矩阵,以拟合邻接矩阵;同时,视社团为文档中主题,隶属度矩阵即可作为 pLSA 模型^[17]中的文档-主题隶属度矩阵,以实现网络拓扑与节点内容的融合. 然后,我们运用节点结构相似度矩阵获取先验信息,基于以下想法:“若两节点具有存在链接且拓扑结构相似的先验信息,则它们可被划分到同一社团”,建模先验信息为一个非负矩阵以调整隶属度矩阵,进而实现先验信息的融入. 最终,我们运用一个平衡因子融合拓扑和内容信息,一个邻域超参控制先验信息,以构建融合节点内容的统一化半监督社团检测模型. 总结来说,本文研究的贡献如下:

(1) 我们提出了一种新型的、统一化的半监督社团检测模型. 该模型不仅融合网络拓扑与节点内容信息,也引入先验信息 must-link 刻画的结构信息.

(2) 在统一模型框架下,获取的单一的节点社团隶属度同时刻画网络拓扑、节点内容和先验三种信息. 模型的鲁棒性得到提升.

(3) 我们通过量化模型中的平衡因子和邻域超参以提升新模型在人工、真实网络上社团检测效果. 与当前流行的社团检测方法相比,新模型识别社团结构的性能更优.

1 相关工作

对本文新模型相关的社团检测算法进行梳理,分别从融合拓扑和内容社团检测、半监督社团检测两方面展开研究.

Ruan 等^[11]基于图聚类提出了 CODICIL 算法,其融合拓扑和内容的方式简单易行,但方法性能依赖于社团结构的清晰程度. Yang 等^[12]基于生成的思想提出了一种联合概率模型 CESNA 融合拓扑和内容,但对拓扑中复杂结构敏感. Wang 等^[10]设计了基于非负矩阵分解的 SCI 算法,学习节点属性-社团隶属度以实现对社团的语义解释,但也受到拓扑结构信息影响. Cao 等^[18]考虑节点度不均衡、节点内容流形等问题,提出了 CLNCCD 模型,发现建模拓扑结构有益于社团检测. 分析上述方法发现,虽然这类方法融合了拓扑和内容,但先验信息的利用仍具有必要性.

Allahverdyan 等^[19]将已知社团信息分配给不同节点,基于半监督图聚类实现社团检测. Ma 等^[20]将 must-link 和 cannot-link 约束融入邻接矩阵并进行分解,以实现半监督社团检测. Yang 等^[15]基于隐空间聚类策略,运用图正则以实现先验信息对社团隶属度的惩罚,构建半监督社团检测模型 NMF_LSE 和 NMF_SYM. Jin 等^[16]分别构建非负矩阵、图正则项以表示先验信息的硬、软约束,提出一种基于非负矩阵分解的链接强化框架(简称 ECD),实现半监督社团检测. 虽然这类方法运用了先验信息,却忽略了内容信息的作用,仍会受到链接存在噪声或是缺失的影响.

2 方法

首先,我们在基于非负矩阵分解的生成框架下构建了融合拓扑和内容的基本模型. 其次,计算节点间的结构相似度,建模先验信息,并将其集成到拓扑信息中,以构建融合节点内容的半监督社团检测模型. 最后,使用梯度下降法实现模型参数的学习.

在本文,我们使用 $G(V, E, F)$ 描述一个无向非加权的属性网络, $V = \{v_1, v_2, \dots, v_n\}$ 表示网络中节点的集合, $E = \{e_1, e_2, \dots, e_m\}$ 表示网络中链接的集合, $F = \{f_1, f_2, \dots, f_l\}$ 表示网络中节点内容特征的集合. 本文中,邻接矩阵 $A \in \mathbf{R}^{n \times n}$ 表示拓扑信息,即当节点 v_i 和 v_j 之间存在链接时, $a_{ij} = 1$, 否则 $a_{ij} = 0$; 使用内容矩阵 $B \in \mathbf{R}^{n \times l}$ 表示节点内容信息,即当节点 v_i 含有第 j 个特征时 $b_{ij} = 1$, 否则 $b_{ij} = 0$.

2.1 基本模型构建

我们使用矩阵 $X \in \mathbf{R}^{n \times k}$ 描述节点社团隶属度分布,其中 x_{ij} 表示节点 v_i 属于第 j 个社团的倾向. 网络中节点之间的链接取决于这两个节点属于同一个社团的概率. 因此, $x_{it}x_{jt}$ 表示社团 t 中节点 v_i 和 v_j 之间生成的链接数量,则整个网络 k 个社团中 v_i 和 v_j 之间的链接期望数可表示为

$$\hat{a}_{ij} = \sum_{t=1}^k x_{it}x_{jt}, \quad i, j = 1, 2, \dots, n. \quad (1)$$

可用式(1)生成期望邻接矩阵 \hat{A} , 即 $\hat{A} = XX^T$, 以拟合邻接矩阵 A . 拓扑子模型的损失函数可写为

$$Loss_T(X) = \|A - \hat{A}\|_F^2 = \|A - XX^T\|_F^2. \quad (2)$$

若网络中的每个节点对应一篇文档,那么社团对应文档集中的主题,节点内容的属性对应文档的单词. 我们借鉴 pLSA 主题模型^[17]的思想构建内容信息子模型. 首先,我们引入一个描述社团内容的矩阵 $H \in \mathbf{R}^{l \times k}$, h_{jt} 表示内容社团 t 包含第 j 个单词的倾向,那么, $x_{it}h_{jt}$ 表示节点 v_i 属于第 t 个内容社团且该内容社团包含第 j 个单词的倾向. 进一步来说,节点 v_i 与第 j 个特征之间的关系可表示为

$$\hat{b}_{ij} = \sum_{t=1}^k x_{it}h_{jt}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, l. \quad (3)$$

可用式(3)生成期望内容矩阵 \hat{B} , 即 $\hat{B} = XH^T$, 以拟合内容矩阵 B . 内容信息子模型的损失函数可写为

$$Loss_C(X, H) = \|B - \hat{B}\|_F^2 = \|B - XH^T\|_F^2. \quad (4)$$

由式(4)可见,我们将社团内容矩阵 H 与社团隶属度矩阵 X 一样作为变量来优化,与 pLSA 相比,不仅得到了节点内容与内容社团之间的关系,而且还具有更低的时间复杂度,为 $O(nkl)$.

至此,结合拓扑子模型(2)和内容信息子模型(4),我们构建了融合网络拓扑和内容信息的统一化基

本模型,其损失函数可写为

$$Loss(\mathbf{X}, \mathbf{H}) = Loss_T + \alpha \cdot Loss_C = \|\mathbf{A} - \mathbf{X}\mathbf{X}^T\|_F^2 + \alpha \cdot \|\mathbf{B} - \mathbf{X}\mathbf{H}^T\|_F^2, \quad (5)$$

式中, α 为平衡因子,控制基本模型(5)中内容信息比重. 在统一化基本模型中,我们视节点内容中挖掘出的内容社团与从拓扑信息中挖掘出的节点社团是一致的. 那么,非负矩阵分解所获得的社团隶属度矩阵 \mathbf{X} 即运用到节点内容. 因此,我们的统一化模型能够同时利用拓扑信息和内容信息联合学习出网络中的社团结构.

2.2 先验信息建模和融合内容半监督社团检测构建

关于先验信息,我们运用网络拓扑中节点之间相似的结构来刻画,具体可用该节点的邻域来进行描述. 因此,网络中节点 v_i 的拓扑结构可表示为

$$D(i) = \{v_j \in V | (v_i, v_j) \in E\} \cup \{v_i\}. \quad (6)$$

启发于人类社会中“熟人交集越大的两个人在同一社会群体中的概率可能也越大”^[21],则节点 v_i 和 v_j 基于拓扑的结构相似度可定义为

$$\sigma_{ij} = \frac{|D(i) \cap D(j)|}{\sqrt{|D(i)| \times |D(j)|}}. \quad (7)$$

使用式(7)可计算网络中任意两个节点之间的结构相似度,记为相似度矩阵 $\Delta = \{\sigma_{ij}\} \in \mathbf{R}^{n \times n}$. 基于以下思想构建先验信息:如果网络中两个节点之间存在链接并且拓扑结构相似度高,我们就有理由认为这两个节点之间存在 must-link,将这两个节点划入同一个社团. 那么,约束矩阵 $\Omega = \{\omega_{ij}\} \in \mathbf{R}^{n \times n}$ 可定义为

$$\omega_{ij} = \begin{cases} 1, & \text{if } a_{ij} = 1 \text{ and } \sigma_{ij} > \varepsilon \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

式中, ε 为控制 must-link 数量的邻域超参. 约束矩阵 Ω 可对应于一个无向图. 为了实现先验信息集成,我们首先计算图 Ω 的连通分量 $P = \{p_1, p_2, \dots, p_q\}$,并认为每个连通分量中的节点之间都满足 must-link 关系,即同一个连通分量的节点应被划入同一个社团. 借鉴 Jin 等所提 ECD 算法^[16]中强约束的构造方法,我们再构建指示矩阵 $\mathbf{C} \in \mathbf{R}^{n \times q}$ 记录每个节点所属的连通分量,以表示先验信息,同时引入一个非负的辅助矩阵 $\mathbf{Y} \in \mathbf{R}^{q \times k}$. 则节点社团隶属度矩阵 \mathbf{X} 可表示为

$$\mathbf{X} = \mathbf{C}\mathbf{Y}. \quad (9)$$

从式(9)可知,若节点 v_i 和 v_j 之间存在 must-link,我们使指示矩阵 \mathbf{C} 的第 i 行和 j 行相等,即 $c_i \cdot = c_j \cdot$,从而, $x_i \cdot = c_i \cdot \mathbf{Y} = c_j \cdot \mathbf{Y} = x_j \cdot$,表示矩阵 \mathbf{X} 的第 i 和 j 行相等,即节点 v_i 和 v_j 有相同的隶属度分布,必被划入同一个社团中. 基于此,我们建模了先验信息.

最终,构建融合节点内容且集成 must-link 先验信息的新模型,其损失函数为

$$Loss_{\text{new}}(\mathbf{Y}, \mathbf{H}) = \|\mathbf{A} - \mathbf{C}\mathbf{Y}\mathbf{Y}^T\mathbf{C}^T\|_F^2 + \alpha \cdot \|\mathbf{B} - \mathbf{C}\mathbf{Y}\mathbf{H}^T\|_F^2. \quad (10)$$

新模型不仅使用先验信息对拓扑信息起到增强的效果,而且还融合了节点内容信息,从而能够更加精确地揭示网络中的社团结构.

3 模型优化

本文基于梯度下降算法学习模型参数. 该模型的参数包括两个参数矩阵 \mathbf{Y} 和 \mathbf{H} . 我们先对公式 $Loss_{\text{new}}(\mathbf{Y}, \mathbf{H})$ 进行迹运算,然后,基于参数矩阵求偏导,进而获取参数更新规则. 损失函数的迹 $L(\mathbf{Y}, \mathbf{H})$ 如下:

$$L(\mathbf{Y}, \mathbf{H}) = \text{tr}(\mathbf{C}\mathbf{Y}\mathbf{Y}^T\mathbf{C}^T\mathbf{C}\mathbf{Y}\mathbf{Y}^T\mathbf{C}^T) + \alpha \cdot \text{tr}(\mathbf{C}\mathbf{Y}\mathbf{H}^T\mathbf{H}\mathbf{Y}^T\mathbf{C}^T) - 2\text{tr}(\mathbf{A}\mathbf{C}\mathbf{Y}\mathbf{Y}^T\mathbf{C}^T) - 2\alpha \cdot \text{tr}(\mathbf{B}\mathbf{H}\mathbf{Y}^T\mathbf{C}^T) + \text{tr}(\mathbf{A}\mathbf{A}^T) + \text{tr}(\mathbf{B}\mathbf{B}^T), \quad (11)$$

式(11)关于参数 \mathbf{Y} 求偏导为

$$\begin{aligned} \frac{\partial L(\mathbf{Y}, \mathbf{H})}{\partial \mathbf{Y}} &= 4\mathbf{C}^T\mathbf{C}\mathbf{Y}\mathbf{Y}^T\mathbf{C}^T\mathbf{C}\mathbf{Y} + 2\alpha\mathbf{C}^T\mathbf{C}\mathbf{Y}\mathbf{H}^T\mathbf{H} - 2\mathbf{C}^T\mathbf{A}^T\mathbf{C}\mathbf{Y} - 2\mathbf{C}^T\mathbf{A}\mathbf{C}\mathbf{Y} - 2\alpha\mathbf{C}^T\mathbf{B}\mathbf{H} = \\ &= 4\mathbf{C}^T\mathbf{C}\mathbf{Y}\mathbf{Y}^T\mathbf{C}^T\mathbf{C}\mathbf{Y} + 2\alpha\mathbf{C}^T\mathbf{C}\mathbf{Y}\mathbf{H}^T\mathbf{H} - 4\mathbf{C}^T\mathbf{A}\mathbf{C}\mathbf{Y} - 2\alpha\mathbf{C}^T\mathbf{B}\mathbf{H}. \end{aligned} \quad (12)$$

根据 Oja 规则^[22],参数 \mathbf{Y} 的更新公式可以写为

$$\mathbf{Y}_{ij}^{\text{new}} = \mathbf{Y}_{ij}^{\text{old}} \cdot \left(\frac{(2\mathbf{C}^T \mathbf{A} \mathbf{C} \mathbf{Y} + \alpha \cdot \mathbf{C}^T \mathbf{B} \mathbf{H})_{ij}}{(2\mathbf{C}^T \mathbf{C} \mathbf{Y} \mathbf{Y}^T \mathbf{C}^T \mathbf{C} \mathbf{Y} + \alpha \cdot \mathbf{C}^T \mathbf{C} \mathbf{Y} \mathbf{H}^T \mathbf{H})_{ij}} \right)^{\frac{1}{4}}.$$

(13)

同理,式(11)关于参数 \mathbf{H} 求偏导为

$$\frac{\partial L(\mathbf{Y}, \mathbf{H})}{\partial \mathbf{H}} = 2\alpha \cdot \mathbf{H} \mathbf{Y}^T \mathbf{C}^T \mathbf{C} \mathbf{Y} - 2\alpha \cdot \mathbf{B}^T \mathbf{C} \mathbf{Y}.$$

(14)

式(14)属于标准非负矩阵分解的优化,同理,依据 Oja 规则,参数 \mathbf{H} 的更新公式可写为

$$\mathbf{H}_{ij}^{\text{new}} = \mathbf{H}_{ij}^{\text{old}} \cdot \left(\frac{(\mathbf{B}^T \mathbf{C} \mathbf{Y})_{ij}}{(\mathbf{H} \mathbf{Y}^T \mathbf{C}^T \mathbf{C} \mathbf{Y})_{ij}} \right)^{\frac{1}{4}}.$$

(15)

在新模型中,我们首先随机初始化 \mathbf{Y} 和 \mathbf{H} . 然后,分别基于更新规则(13)和(15)对 \mathbf{Y} 和 \mathbf{H} 进行迭代更新,直至损失函数(10)收敛为止. 最后,运用式(9)计算出节点社团隶属度,以实现社团检测.

4 实验分析

为了量化新方法检测社团的性能,我们选择一些当前流行的社团检测方法 与 SCDNC 在人工、真实网络上进行性能对比,并运用常用评价方法标准互信息熵 NMI、调整兰德系数 ARI 进行量化.

本文使用的对比算法涉及融合拓扑和内容、半监督社团检测算法两类,具体如表 1 所示. 考虑到算法涉及节点内容,我们使用的人工数据集是基于经典的 GN^[14]、LFR^[25] 基准数据集和结合节点内容生成算法^[26]生成的内容信息构建的新型、具有内容信息的 GN、LFR 基准数据集. 其中,GN、LFR 网络生成算法的参数设置可以参考文献[14–15]. 我们还在真实网络上进行对比实验,真实网络的详细信息如表 2 所示.

表 1 本文实验中的对比方法

Table 1 Comparison methods used in the experiment

Link+Content	SCI ^[10]
	CLNCCD ^[18]
Semi-supervised	NMF_LSE ^[15]
	NMF_SYM ^[15]
	NMF_LSE with Ω
	NMF_SYM with Ω
	ECD ^[16]

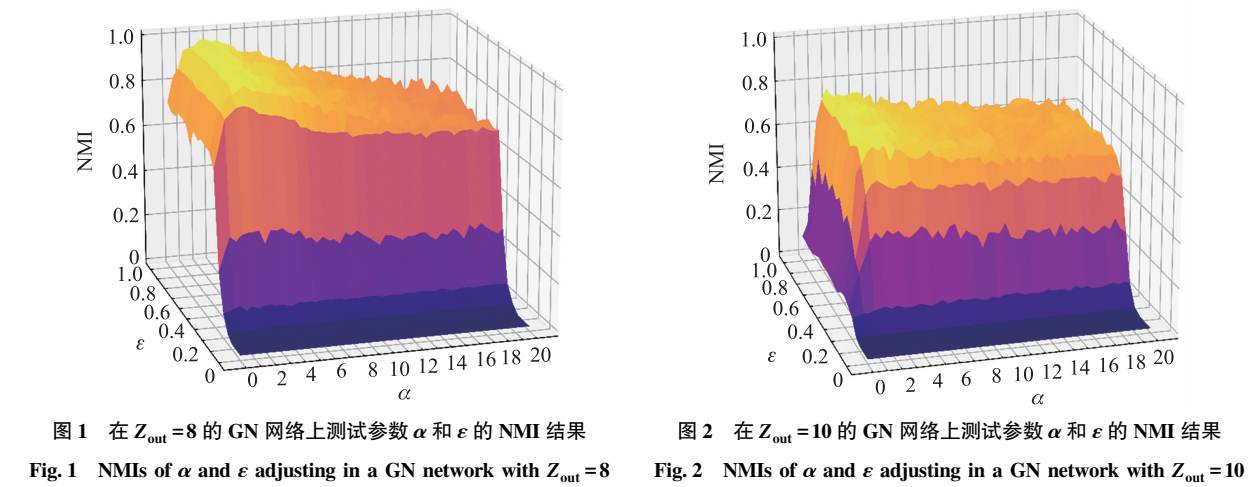
表 2 本文实验中真实网络的详细信息

Table 2 Detailed information of real-world networks used in the experiment^[27]

Dataset	n	m	l	F	k
Cornell	195	283	1 588	0/1	5
Texas	183	276	1 498	0/1	5
Washington	217	366	1 578	0/1	5
Wisconsin	262	459	1 623	0/1	5
Uai2010	3 067	28 308	4 973	0/1	19

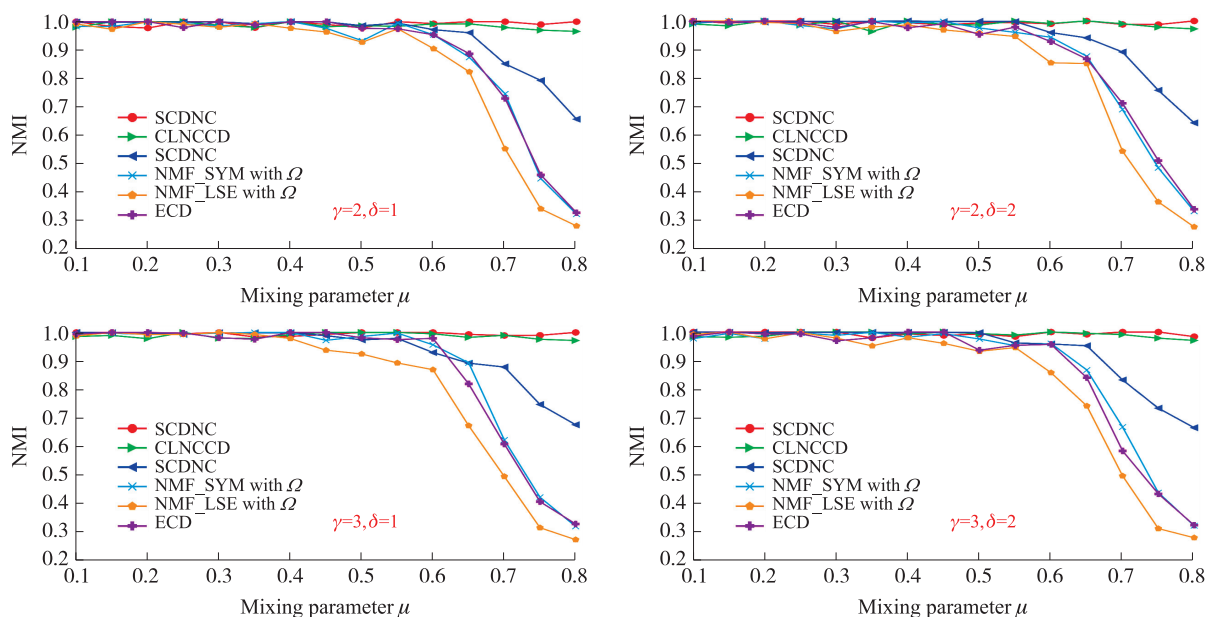
4.1 超参 α 和 ε 敏感度测试

在两个带有节点内容的人工网络 GN 上($Z_{\text{out}}=8$ 和 10,其社团结构非常模糊),我们对超参 α 和 ε 进行测试. 首先为超参选择一个初始化的区域,即 $0<\alpha<20$ 、间隔为 1 和 $0<\varepsilon<1$ 、间隔为 0.05,然后进行遍历搜索. 基于不同的超参组(α, ε),SCDNC 识别社团的精确度 NMI 值曲面如图 1、图 2 所示. 可以发现,当 $(\alpha, \varepsilon) = (3, 0.95)$ 、 $(2, 0.8)$ 时,SCDNC 取得峰值. 综合分析,我们选择 $(\alpha, \varepsilon) = (2, 0.9)$,也将该设置应用于真实网络实验中.



4.2 人工网络上实验分析

我们在网络规模为 $n=1\,000$ 、带有节点内容的 LFR 网络上对比不同方法检测社团的能力. 网络生成算法的混合参数 μ 描述了当前节点与社团外部其他节点之链接所占该节点之度的比例. μ 值越大,网络中社团结构越不清晰. 一般而言,社团检测方法识别该网络中社团的性能呈现下降趋势. 那么,随着 μ 值不断增大,社团检测精度 NMI 曲线下降缓慢的方法更具有竞争力. 在本实验中,我们分别以指数 $\gamma=2,3$ 和 $\delta=1,2$ 的幂律分布生成网络中节点的度和社团的大小,从而得到四组网络. 在每组网络中,我们从 0.1 至 0.8 以 0.05 为步长设计混合参数 μ ,共计 60 个网络作为实验网络数据. 如图 3 所示,不同社团检测方法的基于 NMI 精度曲线随混合参数 μ 值的增大而以不同程度下降. 我们从图中可以较明显地观察到,方法 SCDNC 获得 NMI 曲线下降速度比其他方法更缓慢. 我们还发现,半监督社团检测方法 NMF_SYM with Ω (为 $\alpha=0$ 的 SCDNC)、NMF_LSE with Ω (为 $\alpha=0$ 、基于标准 NMF 的 SCDNC) 和 ECD 获得的精度曲线下降速度快于融合拓扑和内容的社团检测方法 CLNCCD、SCI. 其原因可能是,网络中拓扑与内容具有较强的同质性,以致节点内容具有较强辅助拓扑检测社团的能力. 本小节实验综合分析可得,新方法 SCDNC 融合节点内容、先验信息实现了社团检测性能和鲁棒性的提升.



网络规模 $n=1\,000$, 参数 γ 和 δ 分别表示生成网络算法中度分布和社区规模分布的指数

图 3 我们的方法和对比方法在带有节点内容 LFR 网络上的 NMI 曲线

Fig. 3 Performance comparison of our method and baselines on the LFR benchmark with node contents in terms of NMI

4.3 真实数据集上实验分析

我们还在真实网络上测试 SCDNC 检测社团的性能. 其中,各对比算法中参数设置均为原作者给出的默认设置. 由于 Yang 等[15]未明确设置 NMF_LSE、NMF_SYM 中 must-link 数量. 因此,我们设置 SCDNC 与 NMF_LSE、NMF_SYM 具有相当数量的 must-link. 参考式(8),SCDNC 中 must-link 数量小于且接近网络中链接总数 m ;参考表 2 中真实数据集的 $m/(n(n-1)/2)$ 平均值为 1.48%,可设置 NMF_LSE、NMF_SYM 中 must-link 数量为 $(n(n-1)/2) \times 2\%$,大于且接近 m . 为了检验 SCDNC 融合先验信息的性能,将其与其他半监督方法在不同比例先验信息条件下进行对比. 如图 4 所示,随着 ε 不断增大,SCDNC 的精度明显高于其他方法. 同时,在不同真实网络上,新方法 SCDNC 与其他方法基于评价方法 NMI、ARI 度量的对比结果如表 3 所示. 可以发现,SCDNC 检测不同网络中社团结构,取得最佳或是第二佳的精度值. 尤其运用聚类、社团检测领域中常用的、算法类型无关的评价方法 ARI 评价时,SCDNC 方法仍取得最佳或是第二佳精度. NMF_LSE 在 Uai2010 上取得较高精度,其原因在于使用的 must-link 数量为 SCDNC 所使用的 66 倍之多,先验信息给出社团结构也更清晰. SCDNC 方法识别社团的精度基本高于融合拓扑和内容、半监督类型社团检测方法. 这些实验都验证了,融合节点内容、拓扑的先验信息可进一步提高社团检测性能以及鲁棒性.

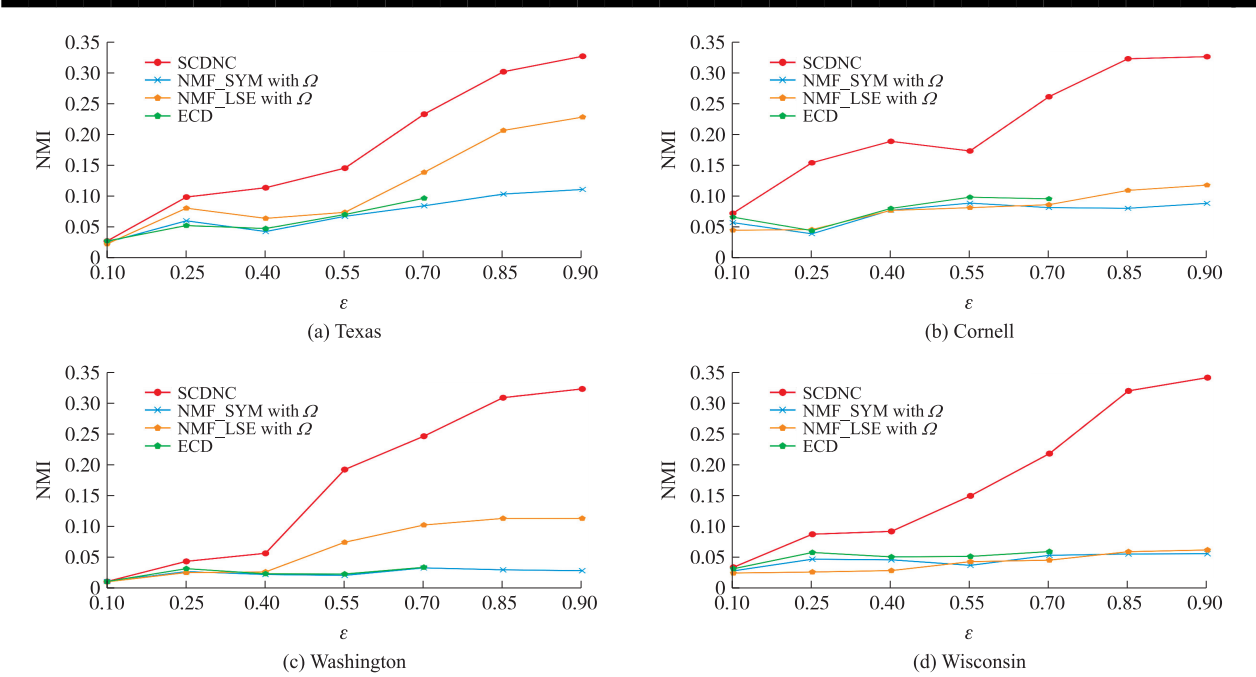


图 4 邻域超参 ϵ 取不同值,不同半监督方法在 Texas(a)、Cornell(b)、Washington(c)和 Wisconsin(d)上的性能对比

Fig. 4 With the different values of the neighborhood hyperparameter ϵ , the performance of different semi-supervised methods on Texas(a), Cornell(b), Washington(c) and Wisconsin(d)

表 3 不同方法社团检测性能对比,加粗、斜体的结果分别指示最佳、第二佳结果

Table 3 Performance comparison of finding communities, the results in BOLD and italic indicate the first-best and the second-best results respectively

Metrics	Methods	Texas	Cornell	Washington	Wisconsin	Uai2010
NMI	SCI	0.1249	0.0680	0.0683	0.1328	0.2340
	CLNCCD	0.2043	0.2168	0.4233	0.2818	0.3685
	NMF_LSE with Ω	0.2260	0.1160	0.1446	0.0863	0.1936
	NMF_SYM with Ω	0.1099	0.0863	0.0362	0.0776	0.2287
	ECD	0.0918	0.0906	0.0379	0.0872	0.2703
	NMF_LSE	0.2625	0.1355	0.2052	0.1762	0.7400
	NMF_SYM	0.1371	0.1052	0.0615	0.1063	0.2826
ARI	SCDNC	0.3236	0.3255	0.4131	0.4843	0.4410
	SCI	0.1793	0.0442	0.1108	0.0655	0.1054
	CLNCCD	0.1935	0.1474	0.3792	0.2222	0.1808
	NMF_LSE with Ω	0.3292	0.0887	0.2403	0.0852	0.0697
	NMF_SYM with Ω	0.2053	0.0344	0.0751	0.0368	0.0875
	ECD	0.1564	0.0522	0.0701	0.0526	0.0899
	NMF_LSE	0.3456	0.1171	0.3041	0.2017	0.5682
	NMF_SYM	0.2682	0.0637	0.1295	0.0929	0.1366
	SCDNC	0.2125	0.2163	0.4272	0.5416	0.2678

4.4 新模型收敛分析

我们在 5 个真实数据集上分析新方法 SCDNC 的模型收敛. 如图 5 所示,SCDNC 的损失函数值随迭代次数增大的变化情况. 在不同规模网络上,SCDNC 需要迭代 300 次左右,其损失函数值的显著变化率小于 10^{-5} ,

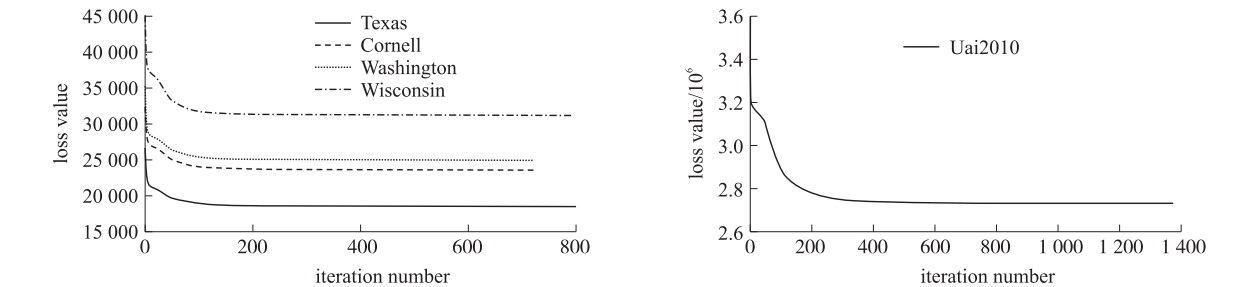


图 5 SCDNC 在 Texas, Cornell, Washington, Wisconsin 和 Uai2010 真实网络上的收敛曲线

Fig. 5 The convergence curve of SCDNC on Texas, Cornell, Washington, Wisconsin and Uai2010, respectively

SCDNC 即可达到收敛.

4.5 新模型的时间复杂度分析

我们还对提出方法进行时间复杂度分析,算法 SCDNC 的主要时空开销包含以下两个过程:一是先验信息的构建过程,另一是目标矩阵 \mathbf{Y} 和 \mathbf{H} 的迭代更新过程. 先验信息的构建过程主要分为三个阶段,相似度矩阵 Δ 的计算、约束矩阵 Ω 的计算以及图 Ω 中连通分量的计算. 这里, n 表示网络中节点数量, m 为网络中链接的数量, k 表示网络中社团的数量, l 为节点内容特征的维度, d_{\max} 为网络中节点的最大度, q 为约束矩阵 Ω 中连通分量的数量. 计算 Δ 、 Ω 以及连通分量的时间复杂度不超过 $O(md_{\max}^2)$ 、 $O(m)$ 和 $O(n)$, 则构建先验信息的时间复杂度为 $O(md_{\max}^2)$. 矩阵 \mathbf{Y} 和 \mathbf{H} 一次更新的时间复杂度分别为 $O(n^2k+nlk)$ 、 $O(nqk+nlk)$. 因此算法迭代一次时间复杂度为 $O(n^2k+nlk)$, 考虑到网络拓扑的稀疏性, 时间复杂度可简化为 $O(mk+nlk)$. 综合以上分析, 假设算法在经历 T 次迭代后达到收敛, 整个算法的时间复杂度将不超过 $O(md_{\max}^2+T(m+n)lk)$, 线性接近于网络规模 n , 具有较低的时间复杂度.

5 结论

我们设计了一种融合节点内容的统一化半监督社团检测模型 SCDNC, 同时结合了网络拓扑、节点内容信息和先验信息. 该模型不仅可借助节点内容减小链接丢失的负面影响, 亦可挖掘网络中结构信息来强化拓扑刻画社团的能力, 具有更优的社团发现性能和鲁棒性. 其关键思想在于基于统一的非负矩阵分解框架充分地整合拓扑、内容和先验信息, 同时运用非负矩阵分解优化以处理 SCDNC 模型的参数学习, 并设计了具有收敛保障的高效更新规则. 实验结果证实了 SCDNC 模型识别网络中社团结构的优越性能.

本文中的真实网络显示拓扑与内容具有较强同质性, 即拓扑和内容刻画一致的社团. 因此, 我们未来将致力于设计能够检测具有非同质拓扑和内容网络中的社团结构的模型. 具体可以通过空间映射等思想尝试实现. 此外, 新模型中社团个数需要预设, 社团个数的自确定也是我们未来需要解决的问题之一.

[参考文献]

- [1] LATORA V, VINCENZO N, GIOVANNI R. Complex networks: principles, methods and applications[M]. Cambridge: Cambridge University Press, 2017.
- [2] RIOLO M A, NEWMAN M E J. Consistency of community structure in complex networks[J]. Physical review E, 2020, 101(5): 052306.
- [3] LESKOVEC J. Large-scale graph representation learning[C]//IEEE International Conference on Big Data. Boston, MA: IEEE, 2017: 4-4.
- [4] 胡云, 张舒, 余侃侃, 等. 基于重叠社区发现的社会网络推荐算法研究[J]. 南京师大学报(自然科学版), 2018, 41(3): 35-41.
- [5] 黄立威, 李彩萍, 张海粟, 等. 一种基于因子图模型的半监督社区发现方法[J]. 自动化学报, 2016, 42(10): 1520-1531.
- [6] 陈俊宇, 周刚, 南煜, 等. 一种半监督的局部扩展式重叠社区发现方法[J]. 计算机研究与发展, 2016, 53(6): 1376-1388.
- [7] JIN D, ZHANG B B, SONG Y, et al. ModMRF: A modularity-based Markov Random Field method for community detection[J]. Neurocomputing, 2020, 405: 218-228.
- [8] NEWMAN M E J, CLAUSET A. Structure and inference in annotated networks[J]. Nature communications, 2016, 7(1): 1-11.
- [9] JIN D, WANG X B, LIU M Q, et al. Identification of generalized semantic communities in large social networks[J]. IEEE transactions on network science and engineering, 2020, 7(4): 2966-2979.
- [10] WANG X, JIN D, CAO X C, et al. Semantic community identification in large attribute networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, AZ: AAAI, 2016: 265-271.
- [11] RUAN Y Y, FUHRY D, PARTHASARATHY S. Efficient community detection in large networks using content and links[C]//Proceedings of the 22nd International Conference on World Wide Web. New York, NY, USA: ACM, 2013: 1089-1098.
- [12] YANG J, MCAULEY J, LESKOVEC J. Community detection in networks with node attributes[C]//IEEE International Conference on Data Mining. Dallas, TX: IEEE, 2013: 1151-1156.
- [13] HE D X, WANG Y Y, CAO J X, et al. A network embedding-enhanced Bayesian model for generalized community detection in

- complex networks[J]. Information sciences,2021,575:306–322.
- [14] GIRVAN M,NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences,2002,99(12):7821–7826.
- [15] YANG L, CAO X C, JIN D, et al. A unified semi-supervised community detection framework using latent space graph regularization[J]. IEEE transactions on cybernetics,2014,45(11):2585–2598.
- [16] HE D X, WANG H C, JIN D, et al. A model framework for the enhancement of community detection in complex networks[J]. Physica A:statistical mechanics and its applications,2016,461:602–612.
- [17] HOFMANN T. Probabilistic latent semantic indexing [C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York,NY:ACM,1999:50–57.
- [18] CAO J X, WANG H C, JIN D, et al. Combination of links and node contents for community discovery using a graph regularization approach[J]. Future generation computer systems,2019,91:361–370.
- [19] ALLAHVERDYAN A E, VER STEEG G, GALSTYAN A. Community detection with and without prior information [J]. Europhysics letters,2010,90(1):18002.
- [20] MA X K, GAO L, YONG X R, et al. Semi-supervised clustering algorithm for community structure detection in complex networks[J]. Physica A:statistical mechanics and its applications,2010,389(1):187–197.
- [21] XU X W, YURUK N, FENG Z, et al. Scan:a structural clustering algorithm for networks [C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose,CA:ACM,2007:824–833.
- [22] CHOI S. Algorithms for orthogonal nonnegative matrix factorization [C]//2008 IEEE International Joint Conference on Neural Networks(IEEE World Congress on Computational Intelligence). Hongkong,China:IEEE,2008:1828–1832.
- [23] LIU H F, WU Z H, LI X L, et al. Constrained nonnegative matrix factorization for image representation[J]. IEEE transactions on pattern analysis and machine intelligence,2011,34(7):1299–1311.
- [24] YEUNG K Y, RUZZO W L. An empirical study on principal component analysis for clustering gene expression data [J]. Bioinformatics,2001,17(9):763–774.
- [25] LANCICHINETTI A, FORTUNATO S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical review E,2009,80(1):016118.
- [26] CAO J X, JIN D, DANG J W. Autoencoder based community detection with adaptive integration of network topology and node contents [C]//International Conference on Knowledge Engineering and Management. Changchun, China: Springer, 2018: 184–196.
- [27] SEN P, NAMATA G, BILGIC M, et al. Collective classification in network data[J]. AI magazine,2008,29(3):93–106.

[责任编辑:杜忆忱]